

Cycles of cooperation and defection in imperfect learning

Tobias Galla*

*Theoretical Physics Group, School of Physics and Astronomy,
The University of Manchester, Manchester M13 9PL, United Kingdom*

When people play a repeated game they usually try to anticipate their opponents' moves based on past observations, and then decide what action to take next. Behavioural economics studies the mechanisms by which strategic decisions are taken in these adaptive learning processes. We here investigate a model of learning the iterated prisoner's dilemma game. Players have the choice between three strategies, always defect (ALLD), always cooperate (ALLC) and tit-for-tat (TFT). The only strict Nash equilibrium in this situation is ALLD. When players learn to play this game convergence to the equilibrium is not guaranteed, for example we find cooperative behaviour if players discount observations in the distant past. When agents use small samples of observed moves to estimate their opponent's strategy the learning process is stochastic, and sustained oscillations between cooperation and defection can emerge. These cycles are similar to those found in stochastic evolutionary processes, but the origin of the noise sustaining the oscillations is different and lies in the imperfect sampling of the opponent's strategy. Based on a systematic expansion technique, we are able to predict the properties of these learning cycles, providing an analytical tool with which the outcome of more general stochastic adaptation processes can be characterised.

I. INTRODUCTION

The mathematical theory of games goes back to von Neumann and Morgenstern [1], and was initially concerned with the study of equilibrium points [2, 3]. The idea that players would be able to compute such equilibria requires severe assumptions, in particular perfect rationality and full knowledge of the game. Additionally each player has to assert that all other players are rational as well. Von Neumann and Morgenstern stress the limitations of their approach explicitly: '*We repeat most emphatically that our theory is thoroughly static. A dynamic theory would unquestionably be more complete and therefore preferable.*' [1].

Since the work of von Neumann and Morgenstern more than 70 years ago, several different routes have been taken to formulate a dynamical theory of games. Evolutionary game theory was launched by Maynard Smith in the 1970s and considers time-dependent dynamics of populations of players [4, 5]. Each individual in the population carries a pure strategy, inherited from its parent(s), and agents then reproduce and pass on their strategies to their offspring, with a reproduction rate depending on the performance in the game. The strategic content of the population evolves, with the concentration of successful strategies increasing over time, and those of less successful strategies being reduced. Evolutionary game theory has been used to model a vast number of phenomena in the social sciences and in economics [6–12].

These applications include in particular the study of the emergence of cooperation and altruism [13]. The evolution of cooperative behaviour under selection pressure constitutes a formidable puzzle. The dynamics of evolution is governed by a fierce competition between individuals, and only those who act in their own interest and who selfishly promote their own evolutionary success at the expense of their competitors should prevail in the long-run. Nevertheless altruism and cooperative behaviour are found in a number of evolved systems, ranging from cooperating genes or cells to cooperating animals or humans in social contexts [14–16]. The question how cooperative behaviour has evolved under strictly competitive and selective dynamics is still unresolved, and has recently been listed as one of the 125 big open problems in science [17].

Our goal here is to address the emergence of cooperation in a third approach to game theory. We focus on adaptive learning processes of a small fixed set of individuals, who interact repeatedly in a game [18–23]. Players observe their opponents' actions and aim to react dynamically by adapting their own strategic propensities, learning from past experience. Such learning models are of particular importance for the understanding of experiments in behavioural game theory, where human subjects play a given game repeatedly under controlled conditions, see e.g. [21–26]. A-priori it is not clear whether adaptation will converge to Nash equilibria. Learning has for example been seen to fail to converge in games with cyclic payoff structures, and complex trajectories including limit cycles, quasiperiodic motion and Hamiltonian chaos have instead been identified

*Electronic address: tobias.galla@manchester.ac.uk

[27–29].

Mathematical models of cooperative behaviour are often based on stylised games played by a small number of interacting individuals, each choosing from a small number of strategies. Such games have been characterised as ‘*mathematical x-ray[s] of crucial features*’ of real-world situations [22]. The most basic setup is the celebrated prisoners’ dilemma, a game in which two players have the choice between cooperation and defection. Defection dominates cooperation in this game, no matter what the other player decides to do, either player will always do better defecting than cooperating. Fully rational players hence end up playing the only equilibrium strategy, defection, and have to put up with the a suboptimal payoff, when they could have scored higher had they both cooperated.

If the prisoners’ dilemma is iterated, more complex behaviour is possible and the space of all strategies grows rapidly as the number of iterations is increased. In order to make progress it is therefore necessary to restrict the mathematical analysis to a subset of this space. We will focus on three strategies: always defect (ALLD), always cooperate (ALLC) and tit-for-tat (TFT). Players using the TFT strategy cooperate in the first iteration and then proceed by playing whatever the opponent played in the previous round. The replicator-mutator dynamics of populations of players engaging in this game have been studied in [30, 31]. ALLD has been identified as the deterministic replicator fixed point, and mutation has been seen to move the attractor toward cooperation. Demographic noise in finite populations can alter the dynamics and can induce coherent evolutionary cycles between defection and cooperation.

As one main result we show that the effects of memory-loss in the learning dynamics are very similar to those of mutation in evolutionary dynamics. While deterministic learning in the absence of memory loss converges to ALLD, this Nash equilibrium is no longer an attractor when players discount observations in the distant past, and a different fixed point, involving all three pure strategies, emerges. Deterministic replicator-type equations are a faithful description of the learning process if and only if a large number of observations of the opponent’s actions is made before players update their own strategic preferences. If, on the contrary, adaptation occurs more frequently and is based only a small sample of observations, the dynamics becomes stochastic. The source of randomness lies in the imperfect sampling of the opponent’s mixed strategy profile. When each player uses a small number of observed actions to estimate the opponent’s mixed strategy, then the estimate will generally be subject to statistical errors. The observed actions were chosen according to the opponent’s mixed strategy profile, but still they are random variables. This source of noise different from the origin of demographic noise in the evolution of finite populations. Nevertheless the effects are similar: as our second main result we show that sustained cycles between cooperation and defection can emerge in stochastic learning, similar to those found in evolutionary scenarios of the iterated prisoner’s dilemma game [30]. We are able to predict the characteristic frequency and power spectra of these cycles analytically as a function of the parameters of the game and the learning dynamics.

II. MODEL

To define the iterated prisoner’s dilemma we will follow the notation of [30]. Assuming that m iterations of the prisoner’s dilemma are played in any one interaction of the two players, and that a complexity cost c is associated with playing TFT the payoff matrix is given by

	ALLC	ALLD	TFT
ALLC	R	S	R
ALLD	T	P	$\frac{T+P(m-1)}{m}$
TFT	$\frac{Rm-c}{m}$	$\frac{S+P(m-1)-c}{m}$	$\frac{Rm-c}{m}$

i.e. a player playing ALLC will for example receive a payoff of R (per round) when meeting another ALLC player, a payoff of S when playing against ALLD and a payoff of R upon encountering TFT. We will denote the payoff matrix elements as a_{ij} , where $i, j = 1, 2, 3$ label the strategies ALLC, ALLD and TFT respectively. Throughout this paper we use $T = 5, R = 3, P = 1, S = 0.1, m = 10, c = 0.8$.

In our model the game is played repeatedly by two players Alice and Bob. We will assume that Alice carries a (time-dependent) mixed strategy profile $\mathbf{x}(t) = (x_1(t), x_2(t), x_3(t))$ and similarly Bob’s mixed strategy profile at t is $\mathbf{y}(t) = (y_1(t), y_2(t), y_3(t))$. We will write $i(t)$ for Alice’s action at time t , and $j(t)$ for Bob’s action, i.e. $i(t), j(t) \in \{\text{ALLC}, \text{ALLD}, \text{TFT}\}$. Following [21–23] each player keeps attractions for each of the pure

strategies. Alice's attractions at time t are labelled by $A_i(t)$ and Bob's attractions by $B_j(t)$. We will again follow [21–23] as well as [27–29] and assume that attractions determine choice probabilities through a logit rule, i.e. that the probabilities for Alice and Bob to play the different pure strategies at time t are given by

$$x_i(t) = \frac{e^{\beta A_i(t)}}{\sum_k e^{\beta A_k(t)}}, \quad y_j(t) = \frac{e^{\beta B_j(t)}}{\sum_k e^{\beta B_k(t)}}. \quad (1)$$

The variable β is a model parameter, and describes the intensity of selection or response sensitivity [23]. For $\beta \rightarrow \infty$ the players strictly choose the pure action with highest attraction, for $\beta = 0$ they play at random. We will here restrict the discussion to models in which both players use the same intensity of selection, generalisation to heterogeneous intensities is straightforward.

A simple re-inforcement learning dynamics is then defined by the following update rules for the attractions

$$\begin{aligned} A_k(t+1) &= (1 - \lambda)A_k(t) + a_{k,j(t)}, \\ B_k(t+1) &= (1 - \lambda)B_k(t) + a_{k,i(t)}. \end{aligned} \quad (2)$$

Alice's attraction A_k is therefore re-inforced by the payoff $a_{k,j(t)}$ she would have received at time t had she played action k , and similarly for Bob. The parameter λ indicates memory loss, observations in the distant past carry a lesser weight than more recent rounds. For $\lambda = 0$ the players have perfect memory of past play, and use the outcome of all past rounds with equal weight to determine their attractions. In particular A_k for example is then the total payoff Alice would have received had she always played action $k \in \{\text{ALLC}, \text{ALLD}, \text{TFT}\}$, given Bob's moves. For $\lambda > 0$ experiences in the past are discounted exponentially. This may happen voluntarily as part of a learning mechanism or simply be due to fading memories and limited mental capacities. We will occasionally refer to λ as a memory-loss rate or discounting factor. We assume that both players learn at identical memory-loss rates, generalisation to heterogeneous learning rules ($\lambda_{\text{Alice}} \neq \lambda_{\text{Bob}}$) is straightforward. Up to relabelling this learning rule is a special case of experience-weighted attraction learning, as discussed in [22, 23]. More general learning dynamics are discussed in the appendix.

The process defined by Eqs. (1,2) is intrinsically stochastic, the actions $i(t)$ and $j(t)$ are drawn from the mixed strategy profiles $\mathbf{x}(t)$ and $\mathbf{y}(t)$ respectively, and accordingly the attractions $A_k(t)$ and $B_k(t)$ are random variables as well. Simple averaging, taking into account that $i(t)$ takes the value $i(t) = \ell$ with probability $x_\ell(t)$ and that $j(t) = \ell$ with probability $y_\ell(t)$, results in the following average attraction update

$$\begin{aligned} A_k(t+1) &= (1 - \lambda)A_k(t) + \sum_{\ell=1}^3 a_{k\ell}y_\ell(t), \\ B_k(t+1) &= (1 - \lambda)B_k(t) + \sum_{\ell=1}^3 a_{k\ell}x_\ell(t). \end{aligned} \quad (3)$$

Limiting dynamics of this type can provide insight into the expected outcome of learning. Deterministic learning has been shown to lead to modified replicator equations in a continuous-time limit [27–29]. Analyses of discrete-time deterministic learning can be found in [32]. The derivation the deterministic dynamics relies on an adiabatic approximation though, it is assumed that strategy updates occur on a much slower time scale than the actual play. In order to perform the update of Eq. (3) Alice has to have full knowledge of Bob's mixed strategy $\mathbf{y}(t)$, and Bob needs to be aware of Alice's strategy $\mathbf{x}(t)$. This will generally be very hard to achieve for the players. Eqs. (3) are therefore only an approximate description of the learning process, and can at best be expected to describe the average behaviour. Describing learning in terms of these deterministic equations is procedurally akin to describing the average behaviour of evolving populations by means of deterministic replicator equations. To understand the nature of the approximation underlying the deterministic limit it is instructive to interpolate between the deterministic average process and the actual stochastic dynamics. We here consider a batch learning process, in which each player samples N actions of their respective opponent, and then updates their attractions. The above 'adiabatic' approximation consists in assuming stationarity of

the mixed strategy profiles between attraction updates. Specifically we introduce the following process

$$\begin{aligned} A_k(\tau + 1) &= (1 - \lambda)A_k(\tau) + \frac{1}{N} \sum_{\alpha=1}^N a_{k,i_\alpha(\tau)}, \\ B_k(\tau + 1) &= (1 - \lambda)B_k(\tau) + \frac{1}{N} \sum_{\alpha=1}^N a_{k,j_\alpha(\tau)}. \end{aligned} \quad (4)$$

The interpretation of these update rules is as follows: at time τ Alice independently selects N actions $i_\alpha(\tau)$ ($\alpha = 1, \dots, N$) following her mixed strategy profile $\mathbf{x}(\tau)$ at that time. I.e. the $\{i_\alpha(\tau)\}$ are independent random variables, and for each α one has $i_\alpha(\tau) = \ell$ with probability $x_\ell(\tau)$. Bob draws his actions $j_\alpha(\tau)$ in a similar manner, using his mixed strategy $\mathbf{y}(\tau)$. These actions represent the moves made by the two players in N successive rounds of the game, the mixed strategies $\mathbf{x}(\tau)$ and $\mathbf{y}(\tau)$ are kept fixed during the course of these rounds. At the end of the batch of N rounds both Alice and Bob update their attractions based on Eq. (4), and then adapt their mixed strategy profiles using Eq. (1) (with t replaced by τ). We have intentionally used the notation τ rather than t to denote time steps of this batch dynamics. One unit of time τ corresponds to N repetitions of the game, i.e. to N units of time t . We will refer to N , the number of observations made in between updates of the attractions, as the batch size, following the language of machine learning [33]. Small batch sizes N correspond to fast adaptation. If $N = 1$ we recover the original dynamics (2) where strategy updates are performed after every single round of the game. Large N on the other hand indicate infrequent adaptation, the limit of infinite batches leads to the deterministic update rule Eq. (3). This limit is based on the assumption that the mixed strategy profiles $\mathbf{x}(\tau)$ and $\mathbf{y}(\tau)$ are stationary during each batch of N repetitions of the game. This assumption will be irrelevant at small batch sizes N , but more severe in the limit of large N . Taking the limit $N \rightarrow \infty$ to derive the deterministic learning rule is analogous to the procedure leading to a description of evolving populations in terms of deterministic replicator equations. In evolutionary systems these descriptions are accurate for populations with an infinite number of individuals. Stochastic corrections cannot be neglected in finite populations, and the resulting noise has been seen to alter the dynamics substantially, see e.g. [30, 31]. Similarly, real-world players do not operate adiabatic learning dynamics, but instead small batch sizes N are probably more appropriate to describe experiments in behavioural economics. It is therefore important to go beyond the deterministic limit of Eq. (3) and to study stochastic effects at finite batch sizes. First steps have been taken in [34], and it is one of the main purposes of this work to apply these ideas to the iterated prisoner's dilemma game.

III. RESULTS AND DISCUSSION

We illustrate the outcome of the continuous-time deterministic learning (see appendix) in Fig. 1. At low memory-loss rates the dynamics is essentially governed by the standard replicator equations, and the system has a single stable fixed point near ALLD, similar to what is reported for low mutation rates in evolutionary systems [30]. As the memory-loss rate is increased ALLD remains a stable attractor, but cyclic attractors around an unstable fixed-point emerge (top right panel of Fig. 1). At even higher memory loss this second fixed point becomes a stable spiral. Provided players do not discount past play too strongly this spiral fixed point is located in the vicinity of the ALLC/TFT edge of the strategy simplex, and we conclude that moderate memory loss may enhance cooperative behaviour. When the memory becomes even shorter the fixed point moves towards the centre of the simplex. In the extreme case of full memory-loss $\lambda = 1$ players ignore the past history beyond the last iteration entirely. Depending on the response sensitivity both players play essentially at random, the three strategies are used with very similar frequencies.

It is interesting to note that the outcome of deterministic learning with memory-loss resembles the behaviour of replicator-mutator dynamics of this game [30]. Discounting past experience in learning and mutation in evolution both promote cooperation when they are moderate in strength. The attractors of learning with quick memory loss on the other hand are similar to those of evolutionary systems in which mutation dominates selection.

We will now move to learning at *finite* batch sizes N . Players are then no longer able to obtain a perfect sample of their opponents' mixed strategy profile before updating their own strategic propensities, and the dynamics becomes stochastic. Results of numerical simulations are shown in Fig. 2. We here focus on a regime in which deterministic learning approaches a fixed point. Stochastic learning at the same discounting rate and intensity of selection results in sustained cycles between cooperation and defection. The amplitude of these cycles is found to scale as $N^{-1/2}$ in the batch size, but the coefficient multiplying $N^{-1/2}$ can be substantial

(see appendix) so that the oscillations can have a significant amplitude. The inset of Fig. 2 confirms that average of several independent runs of the stochastic dynamics is accurately described by the deterministic update rules of Eqs. (3).

The cycling behaviour of the stochastic learning process can be understood as the result of an amplification mechanism, which turns intrinsic white noise into coherent oscillations [35]. The intuitive picture is here as follows: at the memory-loss rate chosen in Fig. 2 the deterministic dynamics spirals into a stable fixed point asymptotically, the relevant eigenvalue of the dynamics is complex. If an instantaneous perturbation were applied to the deterministic system at the fixed point, the dynamics would return to the fixed point following a trajectory of damped oscillations. At finite batch sizes, however, the dynamics is subject to persistent random fluctuations, constantly driving the system away from the fixed point. The combination of this permanent ‘excitation’ and the oscillatory relaxation results in a coherently maintained cyclic pattern.

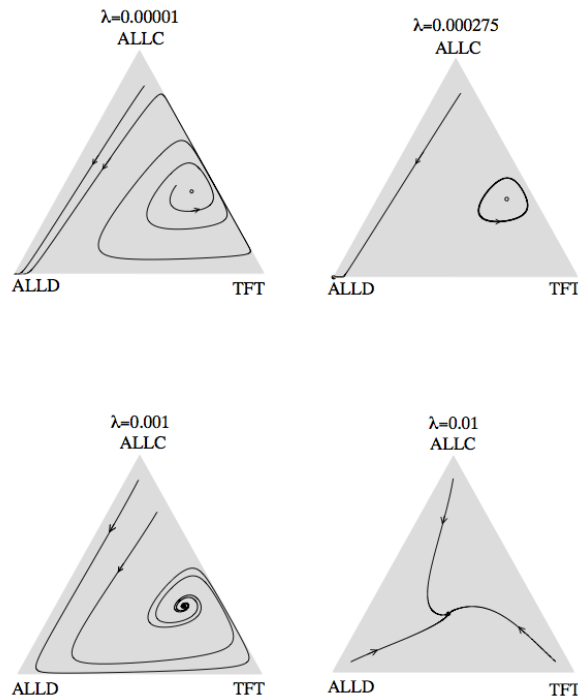


FIG. 1: Illustration of the behaviour of the deterministic continuous-time learning (see appendix) at different memory-loss rates λ . Intensity of selection is $\beta = 0.01$.

Similar noise-induced oscillation phenomena have been observed in various individual-based models of population dynamics, evolutionary game theory, epidemics and biochemical reactions, see e.g. [31, 35–40]. While the mechanism of resonant amplification in stochastic learning is analogous to the one observed in population-based models, the origin of the noise is different. In the individual-based models, large but finite populations are considered. Deterministic mean-field equations can then be derived in the limit of infinite populations. In *finite* populations, the dynamics remains stochastic, due to the random nature of the interactions on the microscopic level. The resulting noise scales with the inverse square root of the system size, and has been termed ‘demographic stochasticity’ [35, 41]. In the learning model the source of the noise is the inaccuracy with which players sample their opponent’s strategy profile at finite batch sizes, and the amplitude of the noise and of the resulting quasi-cycles is proportional to the inverse square root of the batch size N .

We have used a systematic expansion in the inverse batch size to characterise these cycles further (see appendix). These methods are similar to system-size expansions widely used in population-based models [42], even though the expansion parameter in the learning dynamics is the inverse batch size, not the size of the population. Simpler games have been studied with this technique in [34]. These expansion methods are accurate for large, but finite batch sizes. As seen in Fig. 3 the power spectrum of the coherent oscillations can be predicted analytically with great accuracy for moderate and high batch sizes N . The agreement for batch sizes of $N = 10$ is still reasonable, systematic deviations are only found if the number of observations between strategy updates is reduced further.

To characterize the outcome of the stochastic learning process in more details we show the resulting sta-

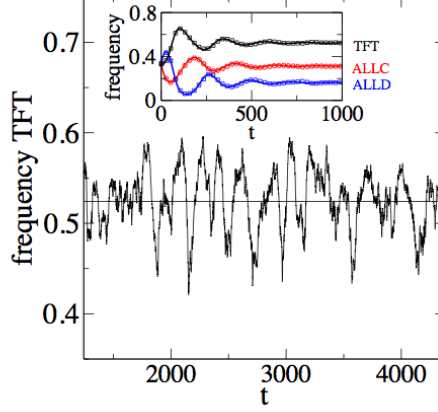


FIG. 2: (Color on-line) Sustained oscillations in the stochastic dynamics. Frequency with which TFT is played by Alice as a function of time at $N = 10$ observations between adaptation events. The horizontal line is the fixed point of deterministic learning. The inset shows the frequencies of ALLC, ALLD and TFT in the initial phase of the dynamics. Solid lines are the outcome of deterministic learning, symbols show data from an average over 100 independent runs of stochastic learning at $N = 10$. Model parameters are $\beta = 0.1, \lambda = 0.01$.

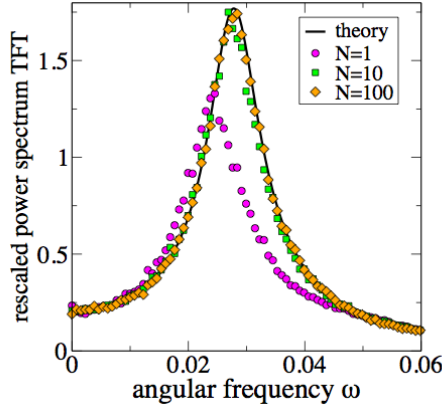


FIG. 3: (Color on-line) Power spectrum of the frequency with which TFT is played. Horizontal axis shows the angular frequency ω , vertical axis the spectrum of fluctuations about the deterministic fixed point. Results from numerical simulations of the stochastic dynamics are shown (markers) along with the curve predicted by the theory in the limit of large, but finite batch size. Power spectra have been re-scaled by the inverse batch size, see appendix. Model parameters are $\beta = 0.1, \lambda = 0.01$. Simulations are averaged over 1000 runs.

tionary distributions in strategy space in Fig. 4. The panels in the upper row correspond to a memory-loss parameter for which the deterministic dynamics has a cyclic attractor. At small batch sizes stochastic learning essentially covers the entire strategy simplex, with the exception of the region near the ALLD/ALLC edge. Surprisingly, the most frequently visited points in strategy space are found along the ALLD/TFT edge, the Nash strategy ALLD is played only very rarely. At larger batch sizes the dynamics concentrates in a region about the deterministic cycle. At fast memory-loss (lower row of Fig. 4) deterministic learning has a fixed point. Again, the stochastic dynamics reaches almost the full strategy space for small batch sizes, but more and more concentration on the deterministic attractor is found as the frequency of adaptation is lowered (i.e. when the batch size is increased). In all cases shown in Fig. 4 the time-average of learning is found near TFT, defection occurs only rarely.

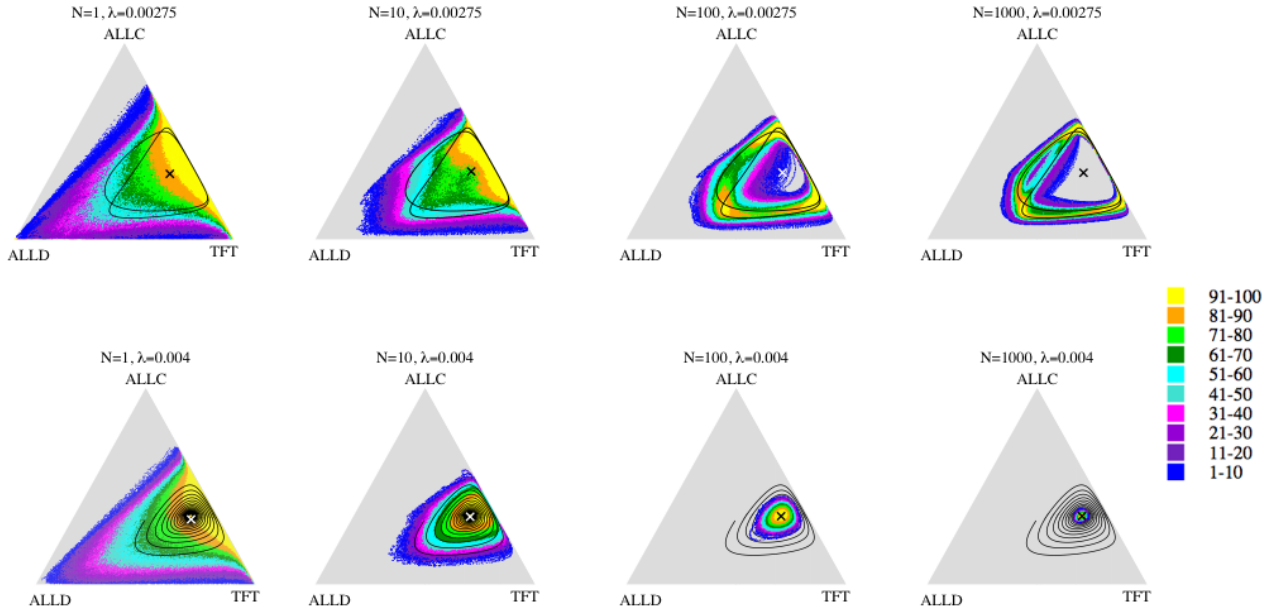


FIG. 4: (Color on-line) Frequencies of visits of the stochastic learning dynamics. Crosses mark the time average of the stochastic dynamics, black lines the trajectory or attractor of the deterministic discrete-time map. Data is obtained from 100 runs of the stochastic process at an intensity of selection $\beta = 0.1$. Colours indicate the frequency with which different regions are visited, a binning of strategy space is performed, and e.g. yellow stands for the 10% most visited bins, orange for the next 10% and so on (see legend). Grey areas are not visited by the dynamics at all in our simulations.

To summarise we have here analysed in detail the learning dynamics of two fixed players interacting in a repeated prisoner's dilemma game. We find that discounting past experience in a deterministic learning produces behaviour very similar to the dynamics found in evolutionary replicator-mutator systems [30]. Memory loss removes the stability of the ALLD fixed point, and leads to attractors near the ALLC/TFT edge of strategy space. In order to go beyond the adiabatic assumption underlying purely deterministic adaptation models, we have also addressed more realistic stochastic learning. Here, players update their strategic propensities more frequently, relying on an imperfect sampling of their opponent's strategy. We then observe persistent stochastic cycles, with a time average concentrated near TFT, paralleling earlier observations in finite evolutionary dynamics [30]. Based on a systematic expansion technique we have characterised these cycles analytically. This method is applicable very generally, and can be used to study the effects of stochasticity in other learning models [18, 22], in machine learning problems and in algorithmic game theory [43]. Cyclic behaviour has been reported in experimental studies of multi-player learning [24], we expect that the techniques we have introduced will be helpful in formulating and calibrating theoretical learning models describing these real-world laboratory experiments.

Acknowledgments

The author would like to thank J. D. Farmer and Y. Sato for useful discussions. This work is supported by a Research Councils UK Fellowship (RCUK reference EP/E500048/1).

-
- [1] von Neumann J. & Morgenstern O. (1944) *Theory of Games and Economic Behavior*. Princeton University Press, 1944.
 - [2] Nash J. (1950) Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36:48–49, 1950.

- [3] Nash J. (1951) Non-cooperative games. *The Annals of Mathematics*, 54:286–295, 1951.
- [4] Price, G. & Maynard Smith, J. (1973) The logic of animal conflict. *Nature*, 246:15, 1973.
- [5] Maynard Smith J. (1998) *Evolution and the theory of games*. Cambridge University, Cambridge UK, 1998.
- [6] Nowak M. A., Komarova N. L. & Niyogi P. (2001) Evolution of universal grammar. *Science*, 291:114–118, 2001.
- [7] Hens T. & Schenk-Hoppe K. (2005) Evolutionary stability of portfolio rules in incomplete markets. *Journal of Mathematical Economics*, 41(1-2):43–66, 2005.
- [8] Myerson R.B. (1991) *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge MA, 1991.
- [9] Gintis H. (2000) *Game Theory Evolving (first edition)*. Princeton University Press, Princeton, NJ, 2000.
- [10] Hofbauer J. & Sigmund K. (1998) *Evolutionary Games and Population Dynamics*. Cambridge Univ. Press, Cambridge 1998.
- [11] Vega-Redondo F. (2003) *Economics and the theory of games*. Cambridge University, Cambridge UK, 2003.
- [12] Nowak M.A. (2006) *Evolutionary Dynamics*. (Harvard University Press, Cambridge, MA).
- [13] Sigmund K. (2010) *The calculus of selfishness*. Princeton University Press, Princeton NJ, 2010.
- [14] Axelrod R. & Hamilton W.D. (1981) The Evolution of Cooperation. *Science* **211** 1390-1396
- [15] Axelrod R. (2006) The Evolution of Cooperation. Revised edition, Perseus Books Group, ISBN 0-465-00564-0
- [16] Nowak M.A. (2006) Five Rules for the Evolution of Cooperation. *Science* **314** 1560-1563
- [17] Pennisi E. (2005) How did cooperative behaviour evolve ? *Science* **309** 93 (2005)
- [18] Fudenberg D., Levine D.K. (1989) *The theory of learning in games*. MIT Press, Cambridge MA, 1989
- [19] Lugosi G. & Cesa-Bianchi N. (2006) *Prediction, learning, and games*. Cambridge University Press, New York, 2006.
- [20] Young H. P. (2004) *Strategic learning and its limits*. Oxford University Press, Oxford UK, 2004.
- [21] Camerer C. F. & Ho T. H. (1999) Experience-weighted attraction learning in normal form games. *Econometrica* **67** (1999) 827
- [22] Camerer C. F. (2003) *Behavioural game theory - Experiments in strategic interaction*. Princeton University Press, Princeton NJ, 2003.
- [23] Chong J. K., Ho T. H. & Camerer C. F. (2007) Self-tuning experience weighted attraction learning in games. *J. Econ. Theory*, 133:177, 2007.
- [24] Semmann D., Krambeck H. J. & Milinski M. (2003) Volunteering leads to rock-paper-scissors dynamics in a public goods game. *Nature*, 425:390, 2003.
- [25] Traulsen A., Semmann D., Sommerfeld R. D., Krambeck H. J. & Milinski M. (2010) Human strategy updating in evolutionary games. *Proceedings of the National Academy of Sciences*, 107(7):2962–2966, 2010.
- [26] Henrich J. et al (Eds). (2004) *Foundations of Human Sociality*. Oxford University Press, Oxford UK, 2004.
- [27] Sato Y., Akiyama E. & Farmer J. D. (2002) Chaos in learning a simple two-person game. *Proc. Nat. Acad. Sci. USA* **99** 4748-4751 (2002)
- [28] Sato, Y. & Crutchfield, J. P. (2003) Coupled replicator equations for the dynamics of learning in multiagent systems. *Phys. Rev. E* **67**, 015206(R)
- [29] Sato Y., Akiyama E. & Crutchfield J. P. (2005) Stability and diversity in collective adaptation. *Physica D* **210** 21-57
- [30] Imhof L. A., Fudenberg D. & Nowak, M. A. (2005) Evolutionary cycles of cooperation and defection. *Proc. Nat. Acad. Sci. USA* **102** 10797-10800 (2005)
- [31] Bladon A. J. , Galla T. & McKane A. J. (2010) Evolutionary dynamics, intrinsic noise and cycles of co-operation. *Phys. Rev. E.* (2010), at press
- [32] Ochea M. (2010) *Essays on nonlinear evolutionary game theory*. PhD thesis, University of Amsterdam, Academic Publishing Services, ISBN 978 90 5170 688 8
- [33] Saad D (Ed.) (1998) *On-line learning in neural networks*. Cambridge University Press, Cambridge UK, 1998
- [34] Galla T. (2009) Intrinsic noise in game dynamical learning. *Phys. Rev. Lett.* **103** 198702
- [35] McKane A. J. & Newman T. J. (2005) Predator-Prey Cycles from Resonant Amplification of Demographic Stochasticity. *Phys. Rev. Lett.* **94**, 218102
- [36] Alonso D., McKane A. J. & Pascual M. (2006) Stochastic amplification in epidemics. *J. R. Soc. Interface* **4** (14) 575582
- [37] Kuske R., Gordillo L. F. & Greenwood P. (2007) Sustained oscillations via coherence resonance in SIR. *Journal of Theoretical Biology* **245** 459-469
- [38] Mobilia M. (2010) Oscillatory Dynamics in Rock-Paper-Scissors Games with Mutations. *Journal of Theoretical Biology* **264** 1-10
- [39] Simoes M., Telo da Gama M. M. & Nunes A. (2008) Stochastic fluctuations in epidemics on networks. *J. R. Soc., Interface* **5** 555-566
- [40] Pineda-Krch M., Blok H. J., Dieckmann U. & Doebeli M. (2007) A tale of two cycles - distinguishing quasi-cycles and limit cycles in finite predator-prey populations. *Oikos* **116** 53-64
- [41] Nisbet R., Gurney W. (1982) *Modelling Fluctuating Populations*. Wiley, New York, 1982
- [42] van Kampen N. G. (1992) *Stochastic Processes in Physics and Chemistry*. Elsevier, New York 1992
- [43] Nisan N., Roughgarden T., Tardos E. & Vazirani V. V. (2007) *Algorithmic game theory*. Cambridge University Press, Cambridge UK, 2007.
- [44] Boland, R. P., Galla T. & McKane, A. J. (2008) How limit cycles and quasi-cycles are related in systems with intrinsic noise. *J. Stat. Mech.* (2008) P09001

- [45] Boland, R. P., Galla T. & McKane, A. J. (2009) Limit cycles, complex Floquet multipliers, and intrinsic noise. Phys. Rev. E 79, 051131

Appendix

1. Deterministic dynamics and modified replicator equations

The limiting deterministic dynamics, obtained for $N \rightarrow \infty$, is given by Eqs. (3). Taking into account Eqs. (1) one can then write the update rule solely in terms of \mathbf{x} and \mathbf{y} and finds the following map [28]

$$\begin{aligned} x_i(t+1) &= \frac{x_i(t)^{1-\lambda} e^{\beta \sum_j a_{ij} y_j(t)}}{\sum_k x_k(t)^{1-\lambda} e^{\beta \sum_j a_{kj} y_j(t)}}, \\ y_j(t+1) &= \frac{y_j(t)^{1-\lambda} e^{\beta \sum_i a_{ji} x_i(t)}}{\sum_k y_k(t)^{1-\lambda} e^{\beta \sum_i a_{ki} x_i(t)}}. \end{aligned} \quad (\text{A5})$$

Taking a continuous-time limit of (3), as discussed in [29], one finds

$$\begin{aligned} \dot{A}_k &= -\lambda A_k + \sum_j a_{kj} y_j, \\ \dot{B}_k &= -\lambda B_k + \sum_i a_{ki} x_i. \end{aligned} \quad (\text{A6})$$

Using Eqs. (1) it is then straightforward to derive deterministic continuous-time evolution equations for the frequencies $\{x_i(t), y_j(t)\}$ with which the pure strategies $i = 1, \dots, S$ are played by the respective players. One finds [28]

$$\begin{aligned} \dot{x}_i &= x_i \beta \left(\sum_k a_{ik} y_k - \sum_{k\ell} x_k a_{k\ell} y_\ell \right) - \lambda x_i \left(\log x_i - \sum_k x_k \log x_k \right), \\ \dot{y}_j &= y_j \beta \left(\sum_k a_{jk} x_k - \sum_{k\ell} y_k a_{k\ell} x_\ell \right) - \lambda y_j \left(\log y_j - \sum_k y_k \log y_k \right). \end{aligned} \quad (\text{A7})$$

These equations are occasionally referred to as the Sato-Crutchfield equations, and it is worth pointing out that they reduce to the standard replicator equations for the case of learning without memory loss ($\lambda = 0$). Furthermore their behaviour is solely determined by the ratio λ/β . If this ratio is fixed then the role of the remaining parameter is merely to set the time scale. It is also easy to verify that the fixed points of Eqs. (A7) coincide with those of Eqs. (A5). The behaviour of these dynamics can be quite intricate, depending on the structure of the underlying game. Sato et al. have for example identified chaotic motion in modified versions of the celebrated rock-paper-scissors game [27–29].

Fig. 1 in the main text has been obtained from a numerical integration of Eqs. (A7), using an Euler-forward scheme. We point out that it is hard to accurately determine the shape of cyclic attractors such as the one in the top-right panel of Fig. 1, even when integrating the dynamics up to large times of up to $5 \cdot 10^6$ and/or at small time stepping ($dt \approx 10^{-3}$). The cycle in Fig. 1 should therefore be understood as an illustration, rather than as a quantitative characterisation of the attractor.

2. Analytical characterisation of stochastic cycles

It is possible to make analytical progress and to compute the spectrum of the oscillations between cooperation and defection analytically in the limit of large, but finite batch sizes N .

We start from the dynamics of Eq. (3),

$$\begin{aligned} A_k(t) &= (1 - \lambda)A_k(t) + \frac{1}{N} \sum_{\alpha=1}^N a_{k,i_\alpha(t)}, \\ B_k(t) &= (1 - \lambda)B_k(t) + \frac{1}{N} \sum_{\alpha=1}^N a_{k,j_\alpha(t)}, \end{aligned} \quad (\text{A8})$$

and note that the expression $\frac{1}{N} \sum_{\alpha=1}^N a_{k,i_\alpha(t)}$ on the right-hand side is a random variable at finite batch sizes N . The same is true for the analogous expression in the update rule for B_k . The mean value of $\frac{1}{N} \sum_{\alpha=1}^N a_{k,i_\alpha(t)}$ is given by $\mu_k(t) = \sum_j a_{kj} y_j(t)$, given that the $j_\alpha(t)$ are drawn from the mixed strategy profile $\mathbf{y}(t)$, i.e. action $\ell \in \{ALLC, ALLD, TFT\}$ occurs with frequency $y_\ell(t)$ on average. Similarly $\frac{1}{N} \sum_{\alpha=1}^N a_{k,j_\alpha(t)}$ has an average of $\nu_k(t) = \sum_i a_{ki} x_i(t)$. Separating off fluctuations, and anticipating their scaling with N , we write

$$\begin{aligned} \frac{1}{N} \sum_{\alpha=1}^N a_{k,i_\alpha(t)} &= \sum_j a_{kj} y_j(t) + \frac{1}{\sqrt{N}} \xi_k(t), \\ \frac{1}{N} \sum_{\alpha=1}^N a_{k,j_\alpha(t)} &= \sum_i a_{ki} x_i(t) + \frac{1}{\sqrt{N}} \eta_k(t). \end{aligned} \quad (\text{A9})$$

By means of the central limit theorem $\xi_k(t)$ and $\eta_k(t)$ can, in the limit of large but finite N , be approximated as Gaussian noise variables of mean zero and with the following correlations

$$\begin{aligned} \langle \xi_k(t) \xi_\ell(t') \rangle &= \delta_{tt'} \sum_j \{ y_j(t) [a_{kj} - \mu_k(t)] [a_{\ell j} - \mu_\ell(t)] \}, \\ \langle \eta_k(t) \eta_\ell(t') \rangle &= \delta_{tt'} \sum_i \{ x_i(t) [a_{ki} - \nu_k(t)] [a_{\ell i} - \nu_\ell(t)] \}, \\ \langle \xi_k(t) \eta_\ell(t') \rangle &= 0. \end{aligned} \quad (\text{A10})$$

Here $\delta_{tt'} = 1$ for $t = t'$ and $\delta_{tt'} = 0$ otherwise. These expressions are obtained for example by writing

$$\xi_k(t) = \sqrt{N} \left[\frac{1}{N} \sum_{\alpha=1}^N a_{k,i_\alpha(t)} - \mu_k(t) \right], \quad (\text{A11})$$

followed by a straightforward evaluation of the above correlators to the appropriate order in $N^{-1/2}$, and taking into account the statistics of the $i_\alpha(t)$.

We can now proceed to insert these expressions into the map (A5) and find

$$\begin{aligned} x_i(t+1) &= \frac{x_i(t)^{1-\lambda} e^{\beta[\sum_j a_{ij} y_j(t) + N^{-1/2} \xi_i(t)]}}{\sum_k x_k(t)^{1-\lambda} e^{\beta[\sum_j a_{kj} y_j(t) + N^{-1/2} \xi_k(t)]}}, \\ y_j(t+1) &= \frac{y_j(t)^{1-\lambda} e^{\beta[\sum_i a_{ji} x_i(t) + N^{-1/2} \eta_j(t)]}}{\sum_k y_k(t)^{1-\lambda} e^{\beta[\sum_j a_{kj} x_j(t) + N^{-1/2} \eta_k(t)]}}. \end{aligned} \quad (\text{A12})$$

Given the presence of the noise terms $\xi_k(t)$ and $\eta_k(t)$, the mixed strategy profiles $\{x_i(t), y_j(t)\}$ will be stochastic variables themselves. The next step is to self-consistently separate deterministic from stochastic contributions, and to derive a closed set of equations describing the evolution of fluctuations about the deterministic limit. To this end we write

$$\begin{aligned} x_i(t) &= \overline{x_i(t)} + \frac{1}{\sqrt{N}} \tilde{x}_i(t), \\ y_j(t) &= \overline{y_j(t)} + \frac{1}{\sqrt{N}} \tilde{y}_j(t), \end{aligned} \quad (\text{A13})$$

where the quantities with overlines represent the deterministic contributions, and quantities with tildes are

stochastic fluctuations. Eqs. (A12) can be written in the form

$$\begin{aligned} x_i(t+1) &= f_i(\mathbf{x}(t), \mathbf{y}(t), \boldsymbol{\xi}(t)), \\ y_j(t+1) &= g_j(\mathbf{x}(t), \mathbf{y}(t), \boldsymbol{\eta}(t)), \end{aligned} \quad (\text{A14})$$

with suitable functions $\{f_i, g_j\}$. One proceeds by substituting (A13) on both sides of Eq. (A14), followed by a systematic expansion in powers of $N^{-1/2}$. To lowest order one finds

$$\begin{aligned} \bar{x}_i(t+1) &= f_i(\bar{\mathbf{x}}(t), \bar{\mathbf{y}}(t), 0), \\ \bar{y}_j(t+1) &= g_j(\bar{\mathbf{x}}(t), \bar{\mathbf{y}}(t), 0), \end{aligned} \quad (\text{A15})$$

i.e. one recovers the deterministic map (A5).

While the calculation up to now applies to any deterministic trajectory, we will from now on restrict the discussion to an asymptotic regime, and assume that the deterministic dynamics has reached a fixed point $\bar{\mathbf{z}}^* = (\bar{\mathbf{x}}^*, \bar{\mathbf{y}}^*)$. This is appropriate in the context of the present investigation, as we are interested in stochastic quasi-cycles about deterministic fixed points. Based on the restriction to deterministic fixed points further analytical progress is relatively straightforward¹.

In next-to-leading order of the expansion in powers of $N^{-1/2}$ one has

$$\begin{aligned} \tilde{x}_i(t+1) &= \sum_k \left(\left. \frac{\partial f_i(\mathbf{x}, \mathbf{y}, \boldsymbol{\xi})}{\partial x_k} \right|_{(\mathbf{x}^*, \mathbf{y}^*, 0)} \tilde{x}_k(t) + \left. \frac{\partial f_i(\mathbf{x}, \mathbf{y}, \boldsymbol{\xi})}{\partial y_k} \right|_{(\bar{\mathbf{x}}^*, \bar{\mathbf{y}}^*, 0)} \tilde{y}_k(t) \right) + \kappa_i(t) \\ \tilde{y}_j(t+1) &= \sum_k \left(\left. \frac{\partial g_j(\mathbf{x}, \mathbf{y}, \boldsymbol{\xi})}{\partial x_k} \right|_{(\mathbf{x}^*, \mathbf{y}^*, 0)} \tilde{x}_k(t) + \left. \frac{\partial g_j(\mathbf{x}, \mathbf{y}, \boldsymbol{\xi})}{\partial y_k} \right|_{(\bar{\mathbf{x}}^*, \bar{\mathbf{y}}^*, 0)} \tilde{y}_k(t) \right) + \rho_j(t) \end{aligned} \quad (\text{A16})$$

where

$$\begin{aligned} \kappa_i(t) &= \beta \left(\bar{x}_i^* \xi_i(t) - \bar{x}_i^* \sum_k \bar{x}_k^* \xi_k(t) \right), \\ \rho_j(t) &= \beta \left(\bar{y}_j^* \eta_j(t) - \bar{y}_j^* \sum_k \bar{y}_k^* \eta_k(t) \right). \end{aligned} \quad (\text{A17})$$

Writing $\mathbf{z} \equiv (z_1, \dots, z_6) = (x_1, x_2, x_3, y_1, y_2, y_3)$, and using the notation $\mathbf{z}(t) = \bar{\mathbf{z}}^* + N^{-1/2} \boldsymbol{\zeta}(t)$ to separate deterministic from stochastic contributions ($\boldsymbol{\zeta} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{y}_1, \tilde{y}_2, \tilde{y}_3)$) one has

$$\boldsymbol{\zeta}(t+1) = \mathbb{J}^* \boldsymbol{\zeta}(t) + \boldsymbol{\varphi}(t), \quad (\text{A18})$$

where \mathbb{J}^* is the 6×6 Jacobian matrix of the deterministic equations (A5), evaluated at the fixed point $\bar{\mathbf{z}}^* = (\bar{\mathbf{x}}^*, \bar{\mathbf{y}}^*)$. The variable $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_6) = (\kappa_1, \kappa_2, \kappa_3, \rho_1, \rho_2, \rho_3)$ represents Gaussian noise, uncorrelated in time, but with cross-correlations between the different components:

$$\langle \varphi_a(t) \varphi_b(t') \rangle = \delta_{tt'} D_{ab}^*. \quad (\text{A19})$$

The elements of \mathbb{D}^* can be expressed in terms of the deterministic variables $\bar{\mathbf{z}}$. More precisely one has, using Eqs. (A17),

$$D_{ij} = \beta^2 \left[\bar{x}_i^* \bar{x}_j^* \langle \xi_i \xi_j \rangle - \bar{x}_i^* \bar{x}_j^* \sum_{k=1}^3 \bar{x}_k^* \langle \xi_i \xi_k \rangle - \bar{x}_j^* \bar{x}_i^* \sum_{k=1}^3 \bar{x}_k^* \langle \xi_j \xi_k \rangle + \bar{x}_i^* \bar{x}_j^* \sum_{k=1}^3 \sum_{\ell=1}^3 \bar{x}_k^* \bar{x}_\ell^* \langle \xi_k \xi_\ell \rangle \right] \quad (\text{A20})$$

¹ A full analytical characterisation of stochastic effects is possible also for periodic attractors of the deterministic dynamics, this has been discussed in the context of chemical reaction systems in [44] and [45]. Such approaches are based on Floquet theory, and we expected that they are applicable also in the learning scenario (with suitable modifications to accommodate the discrete-time dynamics). This is beyond the scope of the work presented in this paper. We point out however that all equations up to (A22) are valid for any deterministic trajectory, provided the fixed point values $\bar{\mathbf{z}}^*$ in the relevant expressions are replaced by their time-dependent counterparts.

and

$$D_{i+3,j+3} = \beta^2 \left[\bar{y}_i^* \bar{y}_j^* \langle \eta_i \eta_j \rangle - \bar{y}_i^* \bar{y}_j^* \sum_{k=1}^3 \bar{y}_k^* \langle \eta_i \eta_k \rangle - \bar{y}_j^* \bar{y}_i^* \sum_{k=1}^3 \bar{y}_k^* \langle \eta_j \eta_k \rangle + \bar{y}_i^* \bar{y}_j^* \sum_{k=1}^3 \sum_{\ell=1}^3 \bar{y}_k^* \bar{y}_\ell^* \langle \eta_k \eta_\ell \rangle \right] \quad (\text{A21})$$

for $i, j \in \{1, 2, 3\}$. The noise variables φ_a with $a \in \{1, 2, 3\}$ are uncorrelated from those with $a \in \{4, 5, 6\}$ so that the matrix \mathbb{D}^* is block diagonal (D_{ab}^* and D_{ba}^* both vanish if $a \in \{1, 2, 3\}$ and $b \in \{4, 5, 6\}$). The covariances of the noise variables $\{\xi_k\}$ and $\{\eta_k\}$ are given by (A10). One further potentially subtle point deserves some attention here. The covariance elements of the noise variables $\{\xi_k\}$ and $\{\eta_k\}$ as given in (A10) depend on the variables $\mathbf{z}(t) = (x_1(t), x_2(t), x(t), y_1(t), y_2(t), y_3(t))$. These in turn have deterministic and stochastic contributions, $\mathbf{z} = \bar{\mathbf{z}}^* + N^{-1/2} \boldsymbol{\zeta}$. Within our expansion in powers of $N^{-1/2}$ it is justified to self-consistently suppress the stochastic contributions $N^{-1/2} \boldsymbol{\zeta}$ to the variables \mathbf{z} in Eq. (A10), as these contributions would not affect results to the order of $N^{-1/2}$ we are working at. For the purposes of Eqs. (A20) and (A21) we therefore use

$$\begin{aligned} \langle \xi_k(t) \xi_\ell(t') \rangle &= \delta_{tt'} \sum_j \{ \bar{y}_j^* [a_{kj} - \bar{\mu}_k^*] [a_{\ell j} - \bar{\mu}_\ell^*] \} \\ \langle \eta_k(t) \eta_\ell(t') \rangle &= \delta_{tt'} \sum_i \{ \bar{x}_i^* [a_{ki} - \bar{\nu}_k^*] [a_{\ell i} - \bar{\nu}_\ell^*] \}, \end{aligned} \quad (\text{A22})$$

where $\bar{\mu}_i^* = \sum_j a_{ij} \bar{y}_j^*$ and $\bar{\nu}_j^* = \sum_i a_{ji} \bar{x}_i^*$.

Starting from the linear equation (A18) we now move to Fourier space and write $\hat{\zeta}_a(\omega)$ for the Fourier transform of $\zeta_a(t)$ and similarly for the noise components φ_a ($a = 1, \dots, 6$). One then has

$$\hat{\zeta}(\omega) = \mathbb{M}^{-1} \hat{\varphi}(\omega), \quad (\text{A23})$$

where $\mathbb{M} = e^{i\omega} \mathbb{I} - \mathbb{J}^*$. The notation \mathbb{I} here indicates the 6×6 identity matrix. The power spectra of the components of $\boldsymbol{\zeta}$ can then be obtained from Eq. (A23), taking into account (A19), i.e. the fact that $\langle \hat{\varphi}_a(\omega) \hat{\varphi}_b(\omega') \rangle = \delta(\omega + \omega') \mathbb{D}_{ab}^*$. One then has

$$P_{aa}(\omega) = \langle |\hat{\zeta}_a(\omega)|^2 \rangle = \sum_{bc} (\mathbb{M}^{-1})_{ab} \mathbb{D}_{bc}^* (\mathbb{M}^{\dagger -1})_{ca}. \quad (\text{A24})$$

The right-hand-side can be evaluated numerically using the explicit form of the Jacobian \mathbb{J}^* and of the noise covariance matrix \mathbb{D}^* . These quantities only depend on the fixed point $\bar{\mathbf{z}}^*$ of the deterministic dynamics, which again can be obtained by numerical iteration of the map (A5), or as a numerical solution of the corresponding fixed point relations.

Power spectra of this type are plotted in Fig. 3. It is important to note that these represent power spectra of the variables $\boldsymbol{\zeta}$, i.e. the pre-factor $1/\sqrt{N}$ in Eq. (A13) have already been scaled out. The theory hence predicts that these re-scaled spectra are independent of the batch size N , which is why the different spectra in Fig. 3 collapse on one curve (with the exception of the $N = 1$ case, at these batch sizes the theory does not apply). To put it in other words: in order to obtain the raw spectra of deviations from the deterministic fixed point the amplitude of the different spectra in Fig. 3 each need to be divided by N . It is then clear that in absolute terms fluctuations are larger for small batch sizes (e.g. $N = 1$) than for larger batches (e.g. $N = 1000$). The amplitude of fluctuations and of the cycles scales as $N^{-1/2}$.

We stress at this point that amplification mechanisms have been studied extensively in population-based models, chemical reaction systems, evolutionary game theory and epidemiology, see for example [31, 35–40]. While the mechanism of amplification is similar to the one discussed here the source of the noise is different. Stochasticity in these population-based models arises when populations are finite, hence the term ‘demographic stochasticity’ [41]. The approach taken in the population models is based on a systematic expansion in the inverse square root of the system size. These techniques are originally due to van Kampen [42]. In the learning system randomness instead comes from imperfect sampling of the opponent’s mixed strategy, and the expansion parameter is the inverse square root of the number of observations made between strategy updates. Similar batch-size expansions have previously been applied to simpler games in [34].

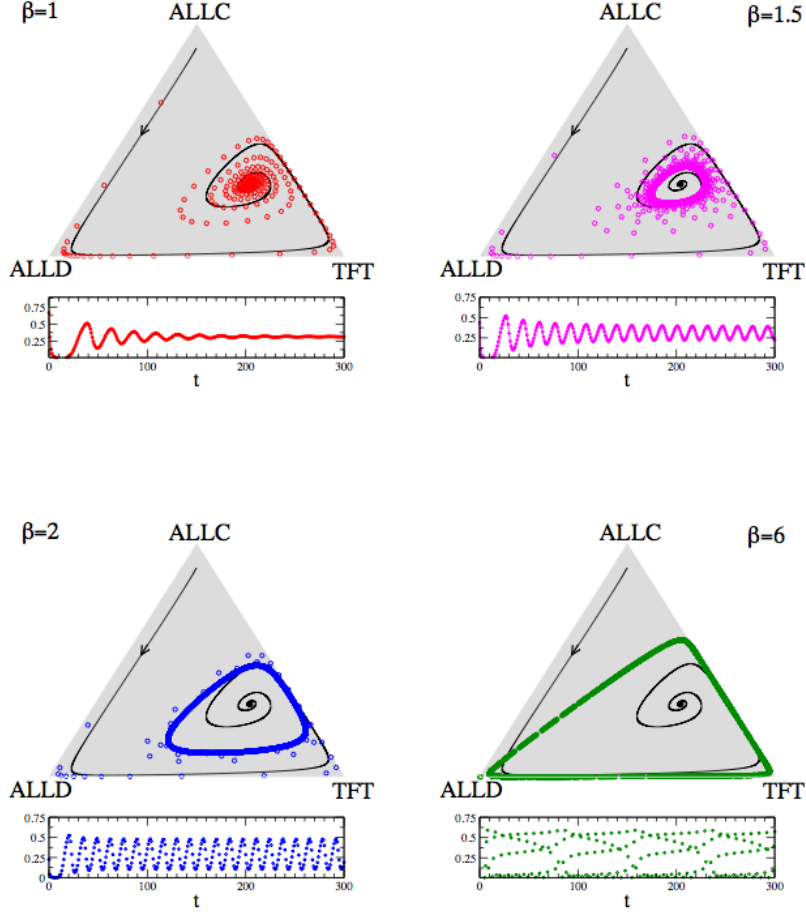


FIG. 5: (Color on-line) Comparison of discrete and continuous-time deterministic learning. We show trajectories of the dynamics at fixed $\lambda/\beta = 0.1$, started from homogeneous initial conditions, $\mathbf{x}(t=0) = \mathbf{y}(t=0)$. The black line in each simplex is obtained from $\beta = 0.01$, and represents the continuous-time limit. The symbols are for $\beta = 1$ (upper left panel), $\beta = 1.5$ (upper right), $\beta = 2$ (lower left) and $\beta = 6$ (lower right). In each panel we show the trajectory in the strategy simplex, as well as the corresponding time series of the propensity, $x_1(t)$ of playing ALLC.

3. Comparison of continuous-time and discrete-time deterministic dynamics

The modified replicator equations, suggested by Sato and Crutchfield [28] are differential equations and as such describe a continuous-time learning process. This approximation is valid for $\beta \ll 1$. The behaviour of discrete-time deterministic learning can however be quite different from this continuous-time limit, as illustrated in Fig. 5. We here fix the ratio λ/β and consider the behaviour at different values of β . For small β the discrete-time maps behaves essentially like the continuous dynamics, and has a stable spiral fixed point. As β is increased however, this fixed point becomes unstable, and a cyclic attractor develops². Further increasing β enlarges the cycle, until its the attractor finally becomes a rather large triangular shaped object as depicted

² While the attractor of the dynamics appears to be a closed cyclic object, it is hard to determine numerically whether the trajectory is actually periodic, as we cannot exclude small drifts. The attractors plotted in the figure may therefore be invariant curves of the map, rather than actual cycles.

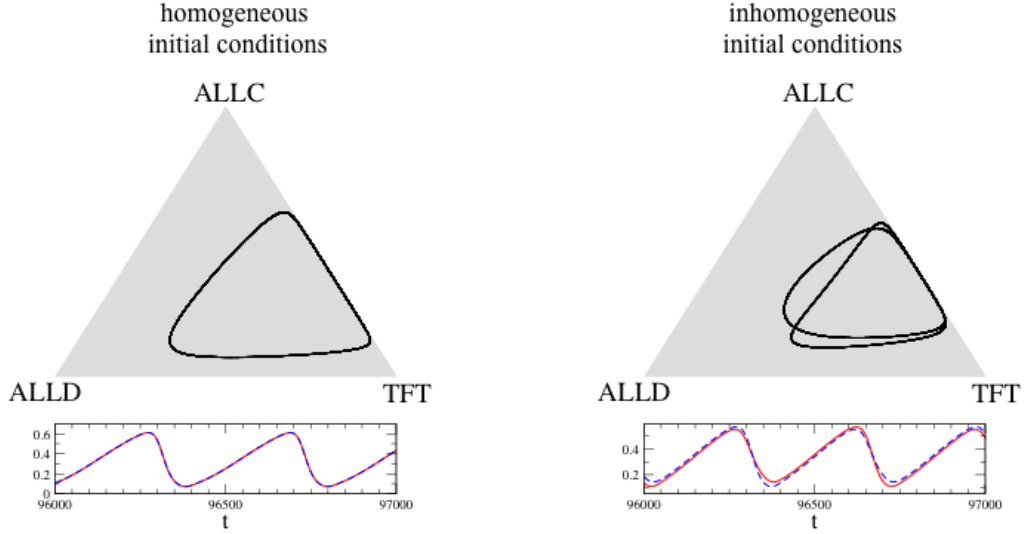


FIG. 6: (Color on-line) Effects of initial conditions. Left: Attractor obtained from running the deterministic map (A5) starting from homogeneous initial conditions. The lower panel shows the ALLC-component of the mixed strategy of each player, the trajectories of both players are identical, $\mathbf{x}(t) = \mathbf{y}(t)$. Right: Attractor obtained from inhomogeneous initial conditions, $\mathbf{x}(t=0) \neq \mathbf{y}(t=0)$. Lower panel shows that the ALLC components of both players are not identical, but that there is a relative shift in time, $\mathbf{x}(t) = \mathbf{y}(t - \Delta t)$. Parameters are $\beta = 0.01$, $\lambda = 0.00275$ in both panels.

in panel d). It is here important to note that even though the attractor set looks smooth, the dynamics does not revolve around the attractor in a continuous motion. We expect that more complicated behaviour, such as chaotic attractors, will in principle be possible, even though we have not observed them for the present game and the present learning dynamics. Other learning rules in similar games have however been shown to admit chaotic motion, see [32].

4. Inhomogeneous initial conditions

The map defined by Eqs. (A5) describes the coupled dynamics between the two players. It is the analogue of a two-population replicator equation in evolutionary dynamics. If started from *homogeneous* initial conditions, $\mathbf{x}(t=0) = \mathbf{y}(t=0)$, the deterministic map will operate in the space in which $\mathbf{x}(t) = \mathbf{y}(t)$, i.e. both players will play identical mixed strategies. This is not generally the case for the stochastic dynamics, as the randomness in the players' decisions will break the symmetry. We find that starting the deterministic map from *inhomogeneous* initial conditions ($\mathbf{x}(t=0) \neq \mathbf{y}(t=0)$) may affect the resulting attractors, see Fig. 6.

5. Effect of selection intensity on stochastic dynamics

The role of the intensity of selection, β , on the stochastic dynamics is illustrated in Fig. 7. The ratio λ/β is the same as in Fig. 4, but with an increased value of β . Comparing Fig. 7 with the upper panels of Fig. 4 shows that an increase of the selection intensity drives the deterministic dynamics to a cycle, and the stochastic dynamics towards the edges of the strategy simplex, especially at small batch sizes N (strong noise). A behaviour not too dissimilar from that of the corresponding evolutionary system emerges, c.f. Fig. 2a and 2c of [30].

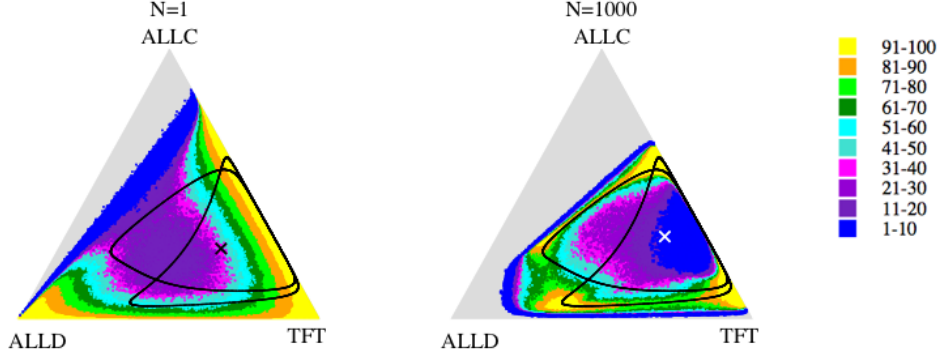


FIG. 7: (Color on-line) Effect of intensity of selection in stochastic dynamics. Figure shows the attractor of deterministic learning (black curves, started from inhomogeneous initial conditions) along with distributions obtained from stochastic learning at $N = 1$ (left) and $N = 1000$ (right) for $\beta = 1$ and $\lambda = 0.04$. The ratio λ/β is thus as in the panels in the upper row of Fig. 4, but with the intensity of selection increased tenfold.

6. Dominance of TFT in stochastic learning at low memory-loss

In [30] it was reported that evolutionary dynamics in the limit of small mutation rates chooses defection at infinite population sizes, but that a finite population of a suitable size can instead choose reciprocity (TFT). An analogous effect is seen in adaptive learning, as illustrated in Fig. 8. We here choose a relatively small memory-loss rate λ , and, as a function of the batch size N , we measure the frequencies with which each of the three pure strategies are played asymptotically in the learning dynamics. While ALLD dominates in the deterministic limit of large N , TFT, i.e. reciprocity is the most frequently used pure strategy in the strongly stochastic case of small batches.

7. Asynchronous updating

The batch dynamics assumes that both players Alice and Bob update their attractions and mixed strategies synchronously once every N rounds of the game. This assumption was made mainly to simplify analytical approaches. In this section we briefly show that asynchronous updating does not alter the picture of coherent stochastic cycles. In Fig. 9 we show time series and power spectra resulting from a learning dynamics in which each player independently updates with probability $1/N$ after each individual round of the game. I.e. one round of the iterated prisoner's dilemma is played, and then for each player it is determined whether or not an update of the player's attractions and mixed strategy occurs (this happens with probability $1/N$), or whether no update is performed (this happens with probability $1 - 1/N$). In each update all rounds of the game since the last update are taken into account. Imposing this dynamics each player performs updates on average every N iterations, but not synchronized with the other player. As seen in the figure coherent cycles are found at finite batches as before.

8. Other learning rules

In this section we briefly consider other learning rules, in particular the experienced-weighted attraction (EWA) learning proposed in [22, 23]. The EWA update follows an algorithm not too dissimilar from the one discussed in the main body of this paper. In particular decisions are based on a logit rule, i.e. we have as before

$$x_i(t) = \frac{e^{\beta A_i(t)}}{\sum_k e^{\beta A_k(t)}}, \quad y_i(t) = \frac{e^{\beta B_i(t)}}{\sum_k e^{\beta B_k(t)}}. \quad (\text{A25})$$

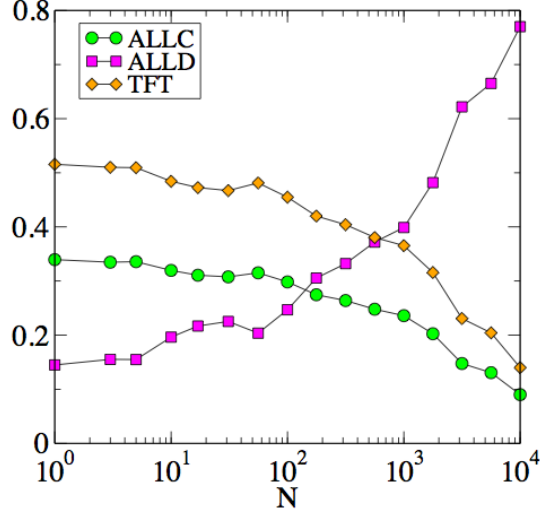


FIG. 8: (Color on-line) Temporal average of the player's mixed strategies in stochastic learning as a function of the batch size. Parameters are fixed at $\beta = 0.01$, $\lambda = 10^{-4}$. At small batch sizes (strong noise) the player's strategies are dominated by TFT, at larger values of N defection is played most frequently. Data is from simulations, run for 100,000 time steps, measurements performed in the second half of this interval, data averaged over 400 samples.

EWA learning uses the following update rule for the attractions $q_i(t)$ and $r_i(t)$:

$$\begin{aligned}
 A_k(t) &= \frac{(1 - \lambda)Z(t-1)A_k(t-1) + [\delta + (1 - \delta)I(i(t), k)] a_{k,j(t)}}{Z(t)} \\
 B_k(t) &= \frac{(1 - \lambda)Z(t-1)B_k(t-1) + [\delta + (1 - \delta)I(j(t), k)] a_{k,i(t)}}{Z(t)}.
 \end{aligned} \tag{A26}$$

Here $i(t)$ is the action taken by player X at round t , and $j(t)$ the action of player y in round t . $I(\cdot, \cdot)$ indicates the Kronecker function, i.e. $I(i, j) = 1$ for $i = j$, and $I(i, j) = 0$ otherwise. The normalisation in the denominator is updated as

$$Z(t) = (1 - \lambda)(1 - \kappa)Z(t-1) + 1. \tag{A27}$$

We note that ϕ in the notation of [22, 23] is equal to $\phi = 1 - \lambda$ in our notation. Eqs. (A26) correspond to on-line learning of batch size $N = 1$. One generalisation to batches of size N is given by

$$\begin{aligned}
 A_k(t+1) &= \frac{(1 - \lambda)Z(t-1)A_k(t-1) + N^{-1} \sum_{\alpha=1}^N [\delta + (1 - \delta)I(i_{\alpha}(t), k)] a_{k,j_{\alpha}(t)}}{Z(t)} \\
 B_k(t+1) &= \frac{(1 - \lambda)Z(t-1)B_k(t-1) + N^{-1} \sum_{\alpha=1}^N [\delta + (1 - \delta)I(j_{\alpha}(t), k)] a_{k,i_{\alpha}(t)}}{Z(t)}.
 \end{aligned} \tag{A28}$$

The parameters β , $\phi = 1 - \lambda$, κ and δ have been fitted to results from real-world experiments on human subjects in [23]. We here focus on the choices $\beta = 1$, $\kappa = 0.75$ and $\phi = 0.8$, these are roughly consistent with values reported in [23], see in particular Table 4 of this reference. It is here important to note that the specific values of these parameter estimates may depend on the detailed experimental protocol, and more importantly on the game under consideration. The purpose of the present section is to show that amplified stochastic oscillations may occur in principle in EWA learning, a more detailed analysis in dependence on model parameters is left for future work.

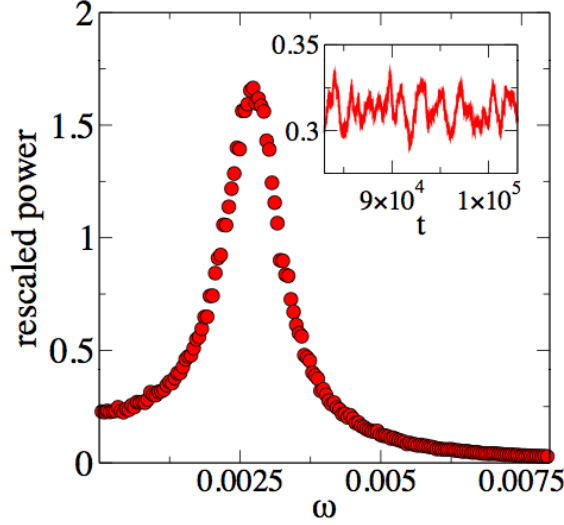


FIG. 9: (Color on-line) Stochastic cycles in asynchronous updating. We show the power spectrum of fluctuations of the propensity to play ALLC (main panel, averaged over 1000 independent runs) as well as a time series from one individual run (inset). Parameters are $\beta = 0.01$, $\lambda = 0.001$, the batch size is $N = 10$.

9. The case $\delta = 1$

We first investigate the case $\delta = 1$, in which all foregone payoffs are re-inforced, i.e. the attractors of all pure strategies are updated, even those of strategies that have not actually been played. Results are shown in Fig. 10, and as seen in the figure the behaviour of the EWA learning model for these parameters is very similar to that of the simplified model of the main body of the paper. The power spectra in the right panel confirm the existence of amplified stochastic oscillations, individual trajectories at different batch sizes are shown in the left-hand panels. The spectra shown in Fig. 10 are obtained from fluctuations about the mean of time series, and that these fluctuations have been re-scaled to take into account the $1/\sqrt{N}$ nature of their magnitude.

10. The case $\delta < 1$

The case $\delta < 1$ is discussed briefly in Fig. 11. Here the strategic choices actually taken are re-inforced with a stronger weight than those which were not played. We find that oscillations persist, provided δ is not too small, the power spectra of fluctuations maintain their maxima at non-zero characteristic frequencies (see Fig. 11). At values of δ smaller than some threshold value (which appears to depend on the other model parameters), no oscillations are found. Near $\delta = 0$ the dynamics may even converge to pure actions. A further more detailed analysis of the EWA model is possible based on the techniques developed in [34] and the present paper. In particular the analytical methods of Sec. A 2 will allow for a detailed study of the regions of parameter space in which cycles between co-operation and reciprocity are to be expected. This will be the topic of future work.

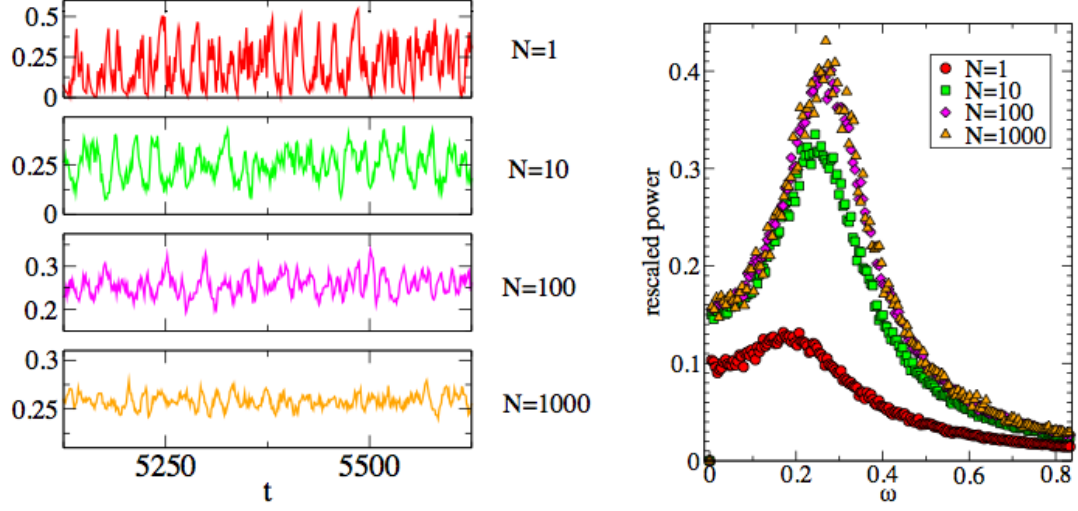


FIG. 10: (Color on-line) Left: Trajectories from single runs of the EWA learning process at $\beta = 1, \kappa = 0.75, \phi = 0.8, \delta = 1$. The curves show the propensity, x_{ALLC} to use ALLC. Right: Corresponding re-scaled power spectra. Simulations are here averaged over at least 100 samples.

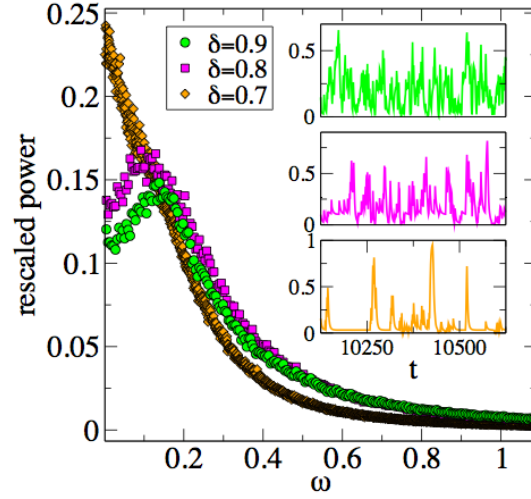


FIG. 11: (Color on-line) The case $\delta < 1$. Other parameters as in Fig. 10. The batch size is $N = 1$. The main panel shows power spectra obtained from time series $x_{ALLC}(t)$ (averaged over 1000 samples), the insets show trajectories $x_{ALLC}(t)$ from individual runs at $\delta = 0.9, 0.8, 0.7$ (from top to bottom).